A.V. Cherkashin, O.V. Anchugova
Ural Federal University named after the first President of Russia B.N. Yeltsin
Yekaterinburg, Russia

DATA ANALYSIS TOOLS IN PYTHON

Abstract: This article provides an overview of data analysis tools and their capabilities in the PYTHON programming language. Assessments of each of these tools are given, their advantages and disadvantages are noted; the tasks they solve and what restrictions they have are described. The article describes what tasks each of the tools solves and how these problems are solved without the tools. The popularity of each of them is estimated based on the popularity of search queries using the Google Trends service for the last 10 years and the results are shown in graphs thanks to which their popularity can be compared. In conclusion there is the description why one tool is more popular than the other.

Keywords: Python, data analysis, machine learning, SciKit-Learn, NumPy, Matplotlib, Pandas.

А.В. Черкашин, О.В. Анчугова
Уральский федеральный университет имени первого Президента России Б.Н. Ельцина
Екатеринбург, Россия

ИНСТРУМЕНТЫ АНАЛИЗА ДАННЫХ НА PYTHON

Аннотация: В данной статье приводится обзор инструментов для анализа данных и их возможностей на языке программирования PYTHON, даны оценки каждого из этих инструментов, отмечены их достоинства и недостатки, а также описаны задачи, которые тот или

иной инструмент позволяет решить и с какими ограничениями они это делают. Описано какие задачи решает каждый из инструментов и как решаются данные задачи без представленных инструментов. Оценена популярность каждого из них на основе популярности поисковых запросов с помощью сервиса Google Trends за последние 10 лет и результаты показаны на графики, благодаря которому можно сравнить их популярность. В заключении делается вывод о том почему один инструмент является популярнее другого.

Ключевые слова: Python, анализ данных, машинное обучение, SciKit-Learn, NumPy, Matplotlib, Pandas.

In recent years, the volume of data has continuously increased. This happens due to the fact that in the contemporary information age, thanks to the opportunities provided by information technology, a large amount of data is collected for each person. The proper example is data from social networks and banking data which make the topic of data analysis relevant.

From the collected data anybody can extract useful information, apply it to personal goals and get the necessary benefits. In some cases, even the standard set of Microsoft Excel tools is suitable. However, for professional and automated data analysis programming languages are used. The most popular in the field of data analysis and statistics are the programming languages Python and R. They both have their advantages and disadvantages. The choice of a programming language depends on the specific situation, the cost of training and the tasks that need to be addressed.

This article focuses on the Python language and the tools it offers to solve data analysis problems. Python has a wide range of applications: web development, game creation, data analysis and many others. For each purpose Python has the tools to make the work easier and data analysis is no exception. In order to find out which specific tools have sufficient functionality and user support to solve the problem, a review and analysis of the main ones will be carried out to assist in mastering this or that technology.

Methods

To assess the popularity of each tool, the Google Trends search queries popularity assessment service is used. With the help of this service,

data for the last 10 years is analyzed, that is over the period from 2009 to 2019, as machine learning and data analysis technologies have been developed for the last 10 years.

To assess the opportunities and identify the strengths and weaknesses of a tool, official documentation and information from open sources are used.

Results

NumPy is an open source library that adds the ability to work with multidimensional arrays and allows them to be processed efficiently. It is the basis for many other libraries for data processing. The main feature is the presence of the array object, which is an array. Graph of popularity is shown in Fig. 1.

Fig. 1 NumPy dynamics of popularity

Pandas is a library with special data structures and operations for managing numeric tables and time series. It allows people to work with two-dimensional and multidimensional tables, allows people to build pivot tables, select columns, use filters by parameters, perform grouping by parameters, run functions and much more. It is high-level, as it is built on top of the NumPy library. It is an alternative to Microsoft Excel, especially useful when analyzing big data. Graph of popularity is shown in the Fig. 2.

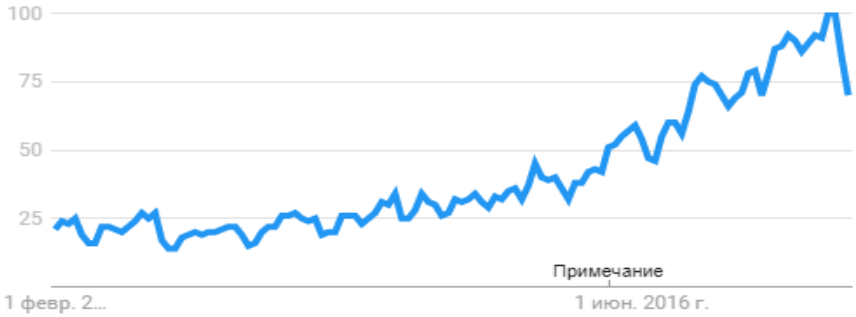

Fig. 2 The dynamics of the popularity of PANDAS

Matplotlib is a library for data visualization. When used in combination with other libraries, it provides features like MATLAB. This library is used to visualize the analyzed data which greatly simplifies the human data perception. Often the resulting images are used as illustrations in publications. Graph of popularity is shown in Fig. 3.

Fig. 3 Dynamics of the popularity of MATPLOTLIB

SciKit-Learn is a machine learning algorithm library used to classify the data studied. It implements all the basic algorithms for machine learning. Moreover, it is relatively easy and it is positioned as a simple library for machine learning but used by many large companies, such as Spotify. Graph of popularity is shown in the Fig. 4.

Fig. 4 Dynamics of the popularity of SciKit-Learn

Discussions

As a result of the study, the data on the popularity of each tool for data analysis over the past 10 years was obtained, as well as on the purpose of each tool and their main capabilities. According to the obtained data, you can compare them according to their capabilities and popularity.

For comparison, compare the graphs for each of the tools.

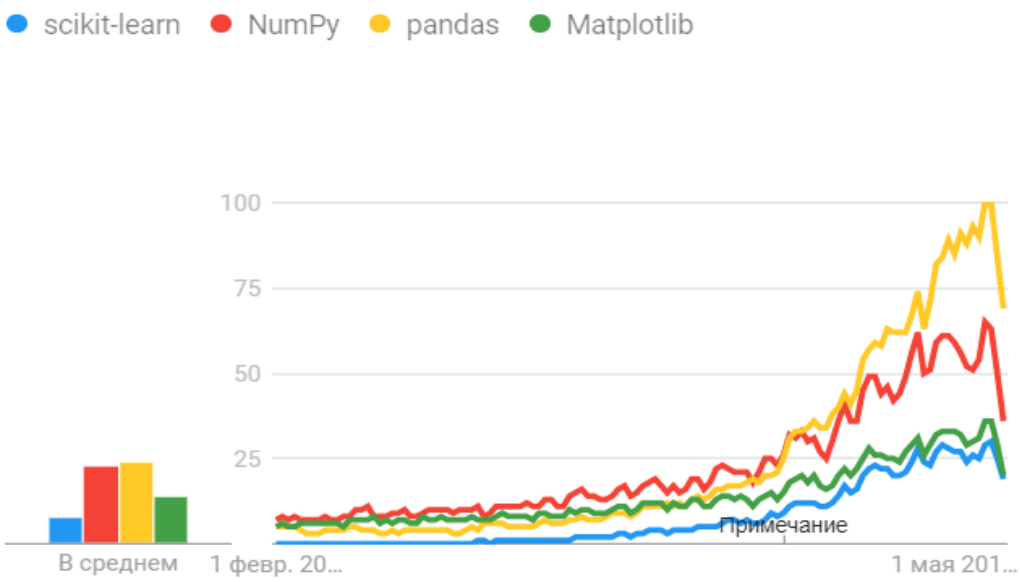

Fig. 5 Comparison of instrument popularity over the past 10 years

From Fig. 5 it can be seen that the tools for data analysis, like the data analysis itself, has become especially popular for the last 2 years. The most popular tool is PANDAS, this is because it is designed to work with tabular data, while the tabular presentation of data is the most convenient and very popular. NumPy also has a high popularity due to the basis for many other libraries. MatplotLib has relatively low popularity due to its

complexity. SciKit-Learn is quite specific and has decent counterparts that provide machine learning algorithms.

According to the results of the study, the tool are not competitors for each other as each of them is designed for a specific range of tasks and occupies its own niche, that is why cannot be replaced, on the contrary, they are used together complementing each other's functionality.

REFERENCES

1. Александр Крот. Введение в машинное обучение с помощью Python и Scikit-Learn [Electronic resource] URL: https://habr.com/company/mlclass/blog/247751/ (08.01.2019).
2. Fogli D., Guida G. Evaluating Quality in Use of Corporate Web Sites: An Empirical Investigation (2018) [Electronic resource] URL: https://dl.acm.org/citation.cfm?doid=3240924.3184646 (05.11.2018).
3. Dumbliauskas V. Grigonis V. Barauskas. A application of google-based data for travel time analysis: Kaunas city case study (2018) [Electronic resource] URL: https://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=5&SID=C16QIBeSgrKKpyX5aA6&page=1&doc=1 (09.12.2018).
4. Peng Lu, Structural effects of participation propensity in online collective actions: Based on big data and Delphi methods (2018) [Electronic resource] URL: https://www.sciencedirect.com/science/article/pii/S0377042718302437 (15.10.2018).
5. Qing Lia, Qianlin Tang, Iotong Chan, Hailong Wei, Yudi Pu, Hongzhen Jiang, Jun Lib, Jian Zhou, Smart manufacturing standardization: Architectures, reference models and standards framework (2018) [Electronic resource] URL: https://www.sciencedirect.com/science/article/pii/S0166361517302075?via%3Dihub (25.10.2018).
6. Ray J., Trovati M., A Survey of Topological Data Analysis (TDA) Methods Implemented in Python(2018) [Electronic resource] URL: https://link.springer.com/chapter/10.1007%2F978-3-319-65636-6_54 (13.11.2018).

7. Ting-Peng Liang, Yu-Hsi Liu, Research Landscape of Business Intelligence and Big Data Analytics: A Bibliometrics Study, Expert Systems With Applications (2018) [Electronic resource] URL: https://www.sciencedirect.com/science/article/pii/S0957417418303099?via%3Dihub (08.10.2018).